36

ANALYZING COMPLEX DATASETS BASED ON THE VARIABILITY FRAMEWORK, DISTRIBUTION ANALYSIS, AND GENERALIZED LINEAR MODELING

Frank Desmet

Introduction

The analysis of human movement in the research domain of embodied music interaction and movement behavior is a complex and challenging phenomenon. The (r)evolution of the development of affordable hardware (computers, sensors, cameras, wireless devices, storage capacities, etc.) to handle large and complex datasets enables researchers to gather large multi-dimensional and mixed datasets with relative ease. The bottleneck in the process of scientific experiments is not the collection of data but rather how to extract useful information from the obtained data based on motivated statistical and scientific thinking. In order to properly analyze data from experiments where humans involved in embodied music interaction is the outcome, one should consider several important issues. This chapter presents three key topics in the field in order to define an appropriate statistical workflow and subsequent analysis to handle complex datasets. It is assumed in this chapter that a suitable experimental design forms the base to obtain a correct dataset. It should be mentioned that formulation of hypotheses and design of experiments are vital clues to obtain reliable data. A good experimental design should be carefully chosen based on a range of factors. Important factors when choosing a suitable design are hypothesis formulation, feasibility, time, cost, ethics, measurement constraints, and what is measured. The design of the experiment is critical for the validity of the results. The aim of this chapter is not to explain hypotheses formulation and experimental design but focuses on strategies to obtain an appropriate analysis of data collected from experiments involving embodied music interaction. Further reading on design of experiments can be found in Goos and Jones (2011). In what follows, the three selected topics are introduced.

The first topic deals with variability. The structure of variability of human movement and interaction in response to music, natural features of the behavior of living organisms, pertains to an interdisciplinary domain. Even in the case of a simple motor task in a controlled lab environment, the potential sources of variability, which can influence the movement/interaction response patterns, are complex and abundant (Stergiou, 2004). Although variability is one of the key concepts of statistical thinking, it is often reduced to a rather naïve concept of "noise" or a "nuisance," which should be "controlled" or "reduced" in order to optimize the central tendency (Gould, 2011). Reporting "statistics" is, even

today, often limited to "mean" values. In the case of production processes or in engineering, noise reduction and/or controlling the variability is precisely what is needed, but when it comes to understanding human behavior (in this case, human movement in response to music), trying to control the variability results in the regulator paradox stated by Weinberg and Weinberg (1979).

In order to understand human responses in a musical context, researchers should avoid playing the role of a "regulator." Commitments and estimates should not be too precise or rigid; otherwise, the flow of information from outside the system is reduced, and this results in losing knowledge of the crucial quality of the system—namely, dynamic adaptation. For example, it is recommended that experiments should not be restricted to the controlled environment of the lab but take place in an ecological setting in order to approximate the real world (Brewer, 2000). Choosing an ecological setting has a great impact on variability of the outcome, as uncontrollable factors are always present in such settings.

Not only is it important to describe the degree of variability (e.g., in terms of contributions of explained and unexplained variability due to deterministic processes), but the global structure of variability should also be taken into account. Rather than focusing on a deterministic methodology by describing separate sources of variability, one should keep in mind that there is a broader context containing other sources of variability, which interact with each other. The theory of embodied music cognition (Leman, 2007) provides a solid base to define a variability framework aiming at a better understanding of the nature of human interaction in a musical context.

The second topic deals with considerations of the distribution of samples. Deviations from the normal distribution occur very often in human responses. *The larger the better* or *The smaller the better* are more often linked with the outcome than *The nominal the better* in experimental designs of human–music research. Especially when an experiment investigates features of entrainment such as synchronization, by means of a repeated measures design, the assumption of the normal distribution is nearly always violated. The most well-known analytical methods such as the *t*-test, *F*-test, linear regression, ANOVA, etc. are based on the assumption of normality and homogeneity of variance and cannot be used to analyze these data.

In the final part of this chapter, generalized linear modeling (GLM) is presented as a valuable tool to handle such complex datasets. Datasets from experiments in systematic musicology are by definition complex (mixed) sets. A complete design exists of pre- and post-survey data and the responses and factor levels from the experiment. In all cases, the models have to be considered as mixed. Continuous, nominal, dichotomous, and ordinal variables can be simultaneously present in the outcome.

Human Movement Variability

Variability in human movement systems is omnipresent and cannot be avoided due to the distinct constraints that define each individual's movement behavior. It is important to notice that not the individual constraints but the confluence of constraints determine the motor action. Rather than considering variability as a source of error, understanding the dynamic structure of variance enables one to define an optimal state of variance. Deviations from this optimal state can result in either rigid (robotic) or unstable (chaotic) movement responses. Variance is often related to instability (i.e., a large variance results in an unstable process). When dealing with human interaction in response to music, this assumption is often violated. For example, an experienced musician can move to music with a high degree of freedom (hence, a large variance), while the overall movement is stable (not chaotic). This depends on the skills of a musician and can be related to expressiveness. Several methods exist to measure and analyze variance. Choosing the best method depends on the underlying properties of a sample and is far from straightforward.

Traditionally, four domains contribute to variability in human movement/interaction response patterns: (1) constraints due to the task, (2) human variability, (3) aggregation, and (4) dynamic

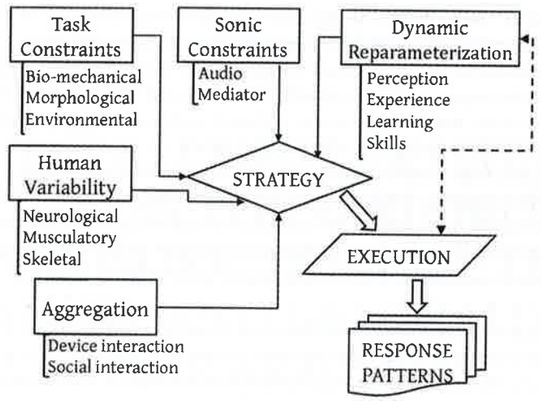

Figure 36.1 Overview of sources of variability of human movement in systematic musicology research. The whole variability framework is embedded in the general musical context.

re-parameterization during the execution of a task. In addition, in the domain of systematic musicology, variability due to sonic constraints is inevitable (Figure 36.1). It is important to consider the figure as an interactive dynamic system (with a time dependency) and not as a sequence of independent static constraints.

In general, the total (observed) variability V_T of a system is the sum of the variability due to nonlinear dynamical processes V_n (including possible interactions) and the variability due to error V_e (Equation 1). The term "error" is somewhat confusing but refers to unexplained variance from unknown sources (noise).

$$V_T = V_n + V_e \tag{1}$$

Each of the five main domains of variability (and their interactions) can contribute to both variability components and are described in the following sections.

Task Constraints

The constraints due to the task can be subdivided in biomechanical, morphological, and environmental conditions and their interactions. For example, if a violin player performs a piece of music, there will be biomechanical constraints due to the way the instrument is played or morphological constraints if one is asked to sit down or stand up (and is able to move more freely) to play, and environmental constraints if, for instance, the task is performed in a recording studio or live on a stage during a festival.

Human Variability

Despite the solid work of Quetelet, who was one of the first scientists (mathematicians) to apply statistical methods to social sciences (Quetelet, 1835), his approach was mainly based on the principle of the "average man," which of course does not exist. Even up until now, examples of this mean approach in peer-reviewed articles can be found (for an overview, see Worthy, 2015).

Especially when looking at the human body, one has to take into consideration that great variability exists among humans and that most of the information can be found in the variability rather than a value of the central tendency. Human variability is a combination of the human system (the body) itself and consists of neurological, skeletal, and muscular limitations of the persons and their interactions. For example, some pianists have the possibility to reach larger scales on the keyboard because they have longer and more flexible fingers (skeletal variability). Other examples are the difference between a healthy trained young adult and an older untrained adult, which have to move on the same piece of music (muscular variability), and the movements on music of people who suffered from a stroke (neurological variability).

Aggregation

Aggregation deals with variability associated with objects or subjects and is divided into human–device interaction and human–human (social) interaction (or a combination of both). The device can be an interactive medium such as a motion sensor, a computer, or a mediator (musical instrument) producing or influencing music, or responding to music. It is well known, for example, that musical performers consider their instrument as an extension of their body. The role of the instrument is that for musicians it is the most natural mediator between subjective experience and physical reality (Nijs, Lesaffre, & Leman, 2013). On the other hand, music is considered to be a social phenomenon, and the way people move in response to music is influenced by the presence of other people. Interactions among performers, between performers and audience, and among listeners are defined as inter-subject relations. The resulting movement variability will be influenced by those factors and may result in aggregation (inter subject-subject, inter subject-device). It means that the movement of a group of individuals will not be a simple addition of the movements of the individuals when moving alone but that intra-subject variability will also contribute to the aggregation variance.

Dynamic Reparameterization

Dynamic reparameterization is related to brain activity while a person interacts with music. It is a continuous combination of brain–body interactions. The difference between research of human body movement and human interaction with music is that nonlinear dynamics should be taken into account in combination with the classical linear approach of movement. The driven force of dynamic re-parameterization is the result of adaptation in combination with reward. The underlying processes can be related to genetic potential and environmental stimuli. It is beyond the scope of this chapter to describe the concept of dynamic reparameterization in detail. It should be mentioned here that the brain does not act as a robotic computer but is an integral part of the human interactive system. Aspects of perception, experience, learning, entrainment, and skills are the main issues that contribute to the variability of human movement/interaction in this domain. Embodied music interaction applied to cognitive neuroscience is a new research field focusing on dynamic reparameterization with potential applications in, for example, the domain of well-being (Lesaffre, 2013) and sports (Buhmann, Desmet, Moens, Van Dyck, & Leman, 2016).

The main difference between this source of variability and the previous three is that there is a continuous neurofeedback while executing a motor task, which results in nonlinear dynamic behavior of human movement related to interaction. If, for example, a subject gets aroused while listening and moving to music, the brain will respond and the neuromuscular system will change during the execution. Factors related to the motor task, human variability, and aggregation are controllable in an experimental design. In contrast, this is rarely the case for the contribution of dynamic reparameterization. Dynamic reparameterization can be either an unconscious or a conscious process. For example, a

subject can deliberately "play around" (e.g., improvising or even deliberately trying to mislead the researcher) while performing a task or is not aware of the fact that her or his movements change due to anticipation or delay.

Sonic Constraints

Variability due to sonic constraints can be divided into two domains. First, several features of the sonic form can contribute to the variability in the response patterns. For example, tempo, energy (loudness), pitch, tonality, and timbre characteristics of the audio may contribute to differences in movement responses. The second domain is that of the mediator, which can be a musical instrument or any type of device that produces the sound. It should be noted that variability due to the mediator is not limited to movements of the performers but that the mediator also contributes to the variability of movements of listeners (participants in an experiment, or an audience). Also, the mediator itself can be a source of variance. An example can be found in interactive sound sculptures that allow people to move freely throughout the installation to create a variety of sounds (Maes, 2009).

Distributional Considerations

Normal Versus Real-Life Distributions

It is common practice to explore a dataset prior to analysis. Procedures such as inspecting the occurrence of missing values, outliers, extreme values, and errors are used to analyze the data. In most cases, a histogram plot is generated (often in combination with P-P and/or Q-Q plots). A histogram provides a visual representation of the shape of the sample distribution. In most cases, the distribution of a population is unknown and has to be estimated from the sample. In the ideal case, the shape is symmetrically bell-curved so that the assumption of a normal distribution can be accepted. In reality, deviations from normality occur very often in human interaction to music. Many popular significance tests such as the *t*-test and ANOVA are of the "parametric" type, meaning they require the data to have certain properties, one of which is "that the population distribution can be considered as normal." Histograms can easily be generated in all kinds of applications (e.g., statistical packages, spreadsheet applications, on-line), but it is usually unclear how the number of bins for the histogram is determined. The number of bins can be determined based on the Friedman-Diaconis, Scott, or Sturges methods (Freedman & Diaconis, 1981; Scott, 1979; Sturges, 1926). This avoids a subjective choice of bins by the researcher, as the number of bins can change the shape of the histogram.

Testing to inspect if the shape of the sample histogram comes from a population with the normal distribution $N(\mu,\sigma)$ is strongly recommended before statistical analysis is performed but often neglected. Different distribution tests are available in most statistical packages (e.g., Kolmogorov-Smirnov, Anderson-Darling, Shapiro-Wilk, Lillifors, and many others; Baghban et al., 2013). If the sample data cannot be assumed to come from a population with a normal distribution, then there are four possible solutions:

1. Consider the distribution as approximately normal distributed.
   Although it seems trivial, this approach is still very popular today and is based on two fundamental theorems of probability: the central limit theorem and the law of large numbers. This "solution" should be avoided and can only be used if there are arguments in favor of the two theorems. This means a test for normality and a sufficiently large sample to obtain sufficient statistical power.

2. Increase the sample size and verify if the deviation from normality is due to too-small samples. Resampling with a larger sample size requires redoing the experiment. This method is often not possible due to cost and time limitations. Furthermore, there is no guarantee that an increased sample size will solve the deviation from normality. Also, a larger sample size can increase precision and power, but when the data have, for example, a heavy-tailed distribution, there will be little improvement of precision and power.
3. Convert non-normal shapes to a normal shape by mathematical transformation.
   Data transformation is sometimes useful but makes interpretation of results more difficult. Some examples of mathematical transformations are linear, logarithmic, square root, and inverse square root. Transformation is not always a guarantee that the transformed data will be normal distributed.
4. Perform tests to find the best corresponding population distribution and choose an appropriate statistical analysis.
   A method to determine the nature of the distribution is based on the shape parameter of a fit of the Weibull distribution. The value of this parameter results in a so-called family of distributions from exponential over normal to the extreme value distribution (Weibull, 1951). The interpretation of the shape parameter can then be used to select the best distribution in the generalized linear model approach explained in the last paragraph of this chapter.

Distributions: Examples

In order to illustrate what real-life distributions from experiments of participants responding to musical stimuli can look like, two examples of typical histogram analysis from two different experiments are presented here (Figure 36.2).

The first example shows the distribution of synchronized counts in an experiment where participants were asked to push a button on the beat of music in different conditions. The histogram shows a left-skewed distribution with oversized zero values and over-dispersion. Furthermore, the dependent variable represents counts. The second example shows the histogram of an experiment where participants had to walk to musical stimuli. The histogram of the "resultant vector length" reveals deviation from normality due to the presence of two overlapping distributions (bimodal): a normal distribution overlapping an extreme value distribution. In this case, two processes were present in the measured dependent variable (Buhmann et al., 2016).

It is clear that analyzing these data from the examples using ANOVA will result in dubious conclusions. In the following paragraph, generalized linear modeling is presented to overcome these problems.

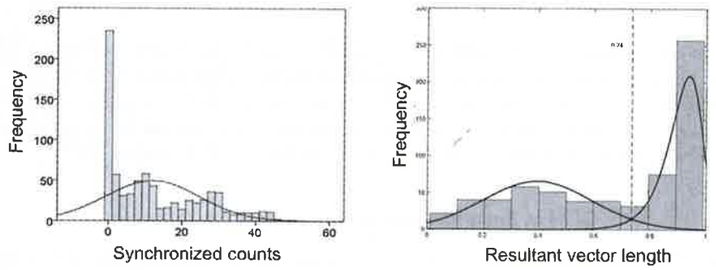

Figure 36.2 Examples of deviation from the normal distribution. Left: counts with oversized zero values, right: bimodal distribution.

Generalized Linear Modeling (GLM)

This section is divided into two parts: first, some basic background information about GLM is presented, and second, an example is given.

Generalized Linear Modeling: An Introduction

In general, models can be considered as abstract, simplified representations of reality. Although much of theoretical statistical inference is based on the assumption that a model is true, one should keep in mind that a model never is. Models involving variability due to unknown random factors have a probabilistic component and are called statistical models. Classical linear models are based on the normal distribution and are well known in statistical analysis and often referred to as general linear models. As mentioned in the previous paragraphs of this chapter, data collected from experiments of human movement and music cannot be analyzed using ANOVA without the risk of making dubious results and interpretations. A solution to overcome this problem is based on the idea of generalized linear models (GLM). Although GLM was introduced in the early 1970s (Nelder & Wedderburn, 1972) and further promoted by Nelder and McCullagh in the 1980s (McCullagh & Nelder, 1989), it took nearly 20 years before this method became more common in use, mainly due to the availability in statistical packages such as SPSS, SAS, and R.

As the name already mentions, GLM is a generalization of the classical linear models. First, these models can involve a variety of distributions belonging to a special family called exponential dispersion models. The exponential family is a broader class of distributions sharing the same density form and including normal, Poisson, gamma, inverse Gaussian, binomial, exponential, and other distributions. Second, the systematic component can be a combination of any data type; and third, a transformation function (link function) is used to link the regression part to the mean of these distributions.

GLM consists of three components:

1. *Random component*: This is the response variable Y and gives the distribution of Y.
2. *Systematic component*: The systematic component specifies the explanatory variables.
3. *Link function*: The link function links the random component with the systematic component.

The advantages of GLM are:

- The flexibility of selecting different distributions.
- No assumption of homogeneity of variance is needed.
- The dependent variable can be of any type: scale, categorical, binomial.
- Complex (mixed) models can be handled.

The main disadvantage of GLM is that it is not straightforward to do. Proper modeling is a difficult task, and calculations are quite complex. The maximum likelihood estimation (MLE) is often a complex mathematical iterative process. Fortunately, nowadays most statistical packages can easily handle the calculations so that this method is available for researchers who do not have a mathematical or statistical background (for further reading, see Dobson & Barnett, 2008; Lindsey, 1997).

GLM in Embodied Music Interaction: An Example

A dataset (N = 792) was obtained in an experiment where three categories of subjects (non-musician, musician, and experienced musician) had to push a button on the beat of four musical excerpts (four BPM levels) in three modalities (auditory-visual, visual, and auditory) and two movement conditions

(no movement and movement). The dependent variable was the number of counts participants synchronized with the musical beat. The general aim of the analysis was to look for significant factors and their interactions.

In order to propose a correct statistical analysis, several problems needed attention. First, the dependent variable is of the interval type (equally spaced ordinal). Histogram analysis based on a number of bins according to the method of Freedman and Diaconis (1981) revealed that the samples are not normal distributed. Also, the occurrence of oversized zero values, over-dispersion (standard deviation greater than the mean), and right-skewed histograms were observed. A one-sample Kolmogorov-Smirnov (Massey Jr, 1951) test for the four excerpts proved that the data are indeed not normal distributed. In addition, a Levene test (Lim & Loh, 1996) revealed also that the homogeneity of variance is violated. Taking these considerations into account, the data could not be analyzed with ANOVA.

A GLM approach based on a Tweedie distribution (Tweedie, 1984) with a log link for a derived dependent variable (the original counts were transformed to percentage of synchronized counts) enabled the data to be analyzed. Tweedie distributions are a family of distributions ranging from normal to gamma distributions, discrete Poisson, and mixed Poisson-gamma distributions. A full factorial two-way analysis was initially performed, excluding non-significant factors and two-way interactions, followed by an analysis of the retained factors and interactions.

Conclusion

This chapter presents three important topics in order to select a suitable statistical method to analyze complex datasets in the field of embodied music interaction.

First, attention is given to several aspects of variability in the domain of human movement in systematic musicology. By considering variability as the key factor of understanding human movement rather than a nuisance or error, a variability framework to map sources of variance is presented. This framework has to be considered as a global system of sources of variance interacting with each other dynamically (time dependency).

Second, considerations about the sampling distribution are given. Techniques to inspect the nature of the sampling distribution based on objective methods are presented. Deviations from the normal distribution are the rule rather than the exception in human movement responses on music. Selecting a motivated statistical strategy to analyze data depends on the nature of the distribution. Although often neglected, this issue is vital for proper statistical analysis in the field.

Finally, generalized linear modeling (GLM) is presented as a method to analyze complex datasets. By using an exponential family of distributions, building up a model where the dependent variable, factors, and covariates can be of any type and a link function related to the chosen distribution in function of the research question can be selected, it is possible to examine mixed datasets. Using GLM it is possible to analyze a combination of survey and experimental data.

Future research will be needed to combine variability and stability. The latter is a real challenge for future research involving embodied music interaction. Methods such as nonlinear dynamics and entropy will be needed to achieve a better understanding of the process of embodied music interaction. These methods enable researchers to consider embodied music interaction as a dynamic process with fluctuations over time.

References

Baghban, A. A., Younespour, S., Jambarsang, S., Yousefi, M., Zayeri, F., & Jalilian, F. A. (2013). How to test normality distribution for a variable: A real example and a simulation study. *Journal of Paramedical Sciences*, *4*(1), 73–77.

Brewer, M. B. (2000). Research design and issues of validity. In H. Reis & C. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 11–26). Cambridge, UK: Cambridge University Press.

Buhmann, J., Desmet, F., Moens, B., Van Dyck, E., & Leman, M. (2016). Spontaneous velocity effect of musical expression on self-paced walking. *PLoS ONE, 11*(5), e0154414.

Dobson, A. J., & Barnett, A. (2008). *An introduction to generalized linear models.* Boca Raton, FL: CRC Press.

Freedman, D., & Diaconis, P. (1981). On the histogram as a density estimator: L 2 theory. *Probability Theory and Related Fields, 57*(4), 453–476.

Goos, P., & Jones, B. (2011). *Optimal design of experiments: A case study approach.* Hoboken, NJ: John Wiley & Sons.

Gould, R. (2011). Variability: One statistician's view. *Statistics Education Journal, 3*(2), 3–15.

Leman, M. (2007). *Embodied music cognition and mediation technology.* Cambridge, MA: MIT Press.

Lesaffre, M. (2013). The power of music and movement to reinforce well-being. In M. Lesaffre & M. Leman (Eds.), *The power of music. Researching musical experiences: A viewpoint from IPEM* (pp. 85–96). Leuven, Belgium: Acco Academic.

Lim, T.-S., & Loh, W.-Y. (1996). A comparison of tests of equality of variances. *Computational Statistics & Data Analysis, 22*(3), 287–301.

Lindsey, J. K. (1997). *Applying generalized linear models.* Berlin, Germany: Springer.

Maes, L. (2009). Sound sculptures and installations as potential new instruments. In S. Carter (Ed.), *Abstracts of the Thirty-Eighth Annual Meeting of the American Musical Instrument Society* (p. 14). Ann Arbor, MI: University of Michigan.

Massey Jr, F. J. (1951). The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American Statistical Association, 46*(253), 68–78.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models.* Boca Raton, FL: CRC Press.

Nelder, J. A., & Wedderburn, W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General), 135*(3), 370–384.

Nijs, L., Lesaffre, M., & Leman, M. (2013). The musical instrument as a natural extension of the musician. In M. Castellengo, H. Genevois, & J.-M. Bardez (Eds.), *Music and its instruments* (pp. 467–484). Editions Delatour France.

Quételet, A. (1835). *Sur l'homme et le developpement de ses facultés ou essai de physique sociale.* Paris, France: Bachelier.

Scott, D. W. (1979). On optimal and data-based histograms. *Biometrika, 66*(3), 605–610.

Stergiou, N. (2004). *Innovative analyses of human movement.* Champaign, IL: Human Kinetics Publishers.

Sturges, H. A. (1926). The choice of a class interval. *Journal of the American Statistical Association, 21*(153), 65–66.

Tweedie, M. C. K. (1984). An index which distinguishes between some important exponential families. In J. K. Ghosh & J. Roy (Eds.), *Statistics: Applications and new directions. Proceedings of the Indian Statistical Institute Golden Jubilee International Conference* (pp. 579–604). Calcutta, India: Indian Statistical Institute.

Weibull, W. (1951). A statistical distribution function of wide applicability. *Journal of Applied Mechanics, 21*, 293–297.

Weinberg, G. M., & Weinberg, D. (1979). *On the design of stable systems.* Hoboken, NJ: John Wiley & Sons.

Worthy, G. (2015). Statistical analysis and reporting: Common errors found during peer review and how to avoid them. *Swiss Medical Weekly, 145*, w14076.